

DOCUMENT RESUME

ED 236 242

TM 830 769

AUTHOR Lutz, William
TITLE What We Know and Don't Know: Needed Research in Writing Assessment.
PUB DATE 12 Apr 83
NOTE 12p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, April 12-14, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Information Analyses (070)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Assessment; Essay Tests; Higher Education; Multiple Choice Tests; *Research Needs; Testing Problems; Test Length; Test Validity; *Writing Evaluation

ABSTRACT

After an extensive review of the available research on large-scale writing assessment, certain issues in writing assessment seem to be unresolved, and still other issues are not supported by adequate research. This paper reviews the basic issues in writing assessment, points out which topics are supported by strong research, and which topics are not supported by research and need to be studied further before any definitive statements can be made about them. Among the topics discussed as needing research are the use of the 45-minute essay as opposed to the 20-minute essay, the use of multiple-choice tests to assess writing, and the effect of item types on writing assessment. Specific topics for further research are listed. (Author)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

What We Know and Don't Know: Needed Research in
Writing Assessment

Paper presented at annual meeting of the National Council
on Measurement in Education
Montreal, Canada
April 12, 1983

I am not a specialist in tests and measurement; I am the head of the English Department at Rutgers University in Camden. My concern with and interest in the assessment of writing stems from my interest in rhetoric and composition. Also, for two years I served as the first director of the New Jersey Basic Skills Assessment Program, a statewide program designed to assess the reading, writing, and mathematics skills of all freshmen entering all New Jersey public colleges and universities. During my term as director, I was responsible for designing, implementing, and administering a test of reading, writing, and mathematics, the results of which were used by New Jersey colleges and universities to place entering freshmen in remedial or non-remedial courses. Of necessity, I quickly learned as much as I could about tests and measurements. Fortunately, I had good instructors from the staff of Educational Testing Service.

It was during this time as director of a statewide testing program that I had occasion to review the research available on writing assessment, a review which I continue even today. I have, at present, a bibliography of some 150 items on writing

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- ☐ This document has been reproduced as received from the person or organization originating it.
☒ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

W. Lutz

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

assessment. But since I knew little about writing assessment, when I started the New Jersey testing program, I asked questions, questions which I thought were obvious, even basic, to constructing any test of writing assessment. Much to my surprise, I discovered that there were few answers available based on research.

My starting point in research on writing assessment, and still the best starting point even now, was The Measurement of Writing Ability by Fred Godshalk and his colleagues at ETS (College Entrance Examination Board, 1966). Any review of research in writing assessment must begin with this important study. While Godshalk does answer some questions, he also raises many questions for future study, questions which have not yet been answered through research.

Godshalk concludes that "when objective questions specifically designed to measure writing skills are evaluated against a reliable criterion of writing skills, they prove to be highly valid." (p. 40) Godshalk goes on to stress that the item types used in the multiple-choice test must be "well designed," a term which he does not specifically define.

Godshalk also concludes that "the reliability of essay scores is primarily a function of the number of different essays and the number of different readings included." The procedure Godshalk used to score the essays was general impression; that is, readers were asked simply to rank the essays according to three scores: 1 for an inferior paper,

2 for an average paper, and 3 for a superior paper. Each paper was read by five readers, the scores added together, and the total correlated with a score on a multiple-choice test. Godshalk's method of reading the essays is the beginning of the procedure which today is called holistic scoring. (Please note that general impression scoring is not holistic scoring, nor is holistic scoring the same as general impression scoring.)

Godshalk is careful to stress in his study that the students were writing under artificial conditions and constraints. Even though each student wrote five essays, they were written over a relatively short period of time. As Godshalk notes:

As in all essays written under test conditions, these place a premium on fluency and ability to write correctly and with some style in a first draft. In actual life situations the writer is seldom under such sharp limitations. He can write and revise. He can consult a dictionary or thesaurus. He can even ask his wife or sweetheart to edit a first draft for him. And he will often have to spend long hours outlining and developing an extensive treatment of a truly complex topic. These samples provide as valid measures of writing under test conditions as any that can be obtained in a similar period of time. It would be interesting to collect samples of writing

completed over a period of several months under much freer conditions and see how scores based on these samples would be related to the measures we have developed. (p. 41)

Godshalk was concerned with the practical as well as the theoretical. He was concerned with designing a method for assessing writing that would not only be reliable and valid but efficient as well. His study concludes that "the most efficient predictor of a reliable direct measure of writing ability is one which includes essay questions . . . in combination with objective questions." (p. 41) Nor was Godshalk unaware of the political implications of testing. "The addition to prediction from the 20-minute essay is real but small . . . It is doubtful that the slight increase in validity alone can justify the increased cost. Rather, the advantage has to be assessed in terms of the model the essay provides for students and teachers." (p. 41) Including an essay says to the student that "skill in actual writing is an important outcome of instruction. It says to the teacher that the ability to answer multiple-choice questions, unless accompanied by the ability to compose answers to essay questions, is not sufficient evidence of effective teaching. (p. 41)

Godshalk's conclusion, simply stated, is that writing ability can be efficiently and effectively assessed by using

a 20-minute essay and a well-designed multiple-choice test, and combining the results of each into one score. This is the procedure used on the New Jersey College Basic Skills Placement Test.

I would like, however, to raise some questions at this point, questions which were not answered by the Godshalk study and which have not as yet been answered in any research that I have been able to locate.

1. Does a 45-minute essay give more information on writing ability than a 20-minute essay?
If so, what kind of information does the 45-minute essay provide that the 20-minute essay does not?
2. Do weak writers perform more poorly on multiple-choice tests than strong writers? Or do strong writers perform poorly on multiple-choice tests?
3. Does it make any difference what item types are used on multiple-choice tests to assess writing?
4. To what extent can multiple-choice tests assess writing? That is, do multiple-choice tests only assess surface or superficial aspects of writing, or can they assess more complicated or sophisticated aspects of writing? If so, which aspects?
5. Is the quality of writing produced by a student under test conditions different from that

produced under non-test conditions? If it is, how different is that quality of writing, and what does that difference say about the validity of current tests of writing?

These are but a few of the questions for which I can find little information in the available literature on the assessment of writing. Indeed, I can find fewer than 200 items of research on the assessment of writing, a remarkably small number of reports and studies on a topic so complicated and important. And many of these studies are incomplete, flawed, and simply badly done. In short, when we turn to the research literature for answers to our questions, we find little to help us.

For example, it is a basic tenet of faith among teachers of writing that any writing produced in a test situation is artificial and not truly representative of the quality of writing which a student can really produce if only given the time and a less artificial situation. Yet based upon my experience as a Director of Freshman English for seven years and a teacher of writing for even more years, I have found very little variation between the prose a student produces in a test essay and the prose produced in a writing course. Now it may be that both situations are artificial and therefore a student produces the same prose in both situations. If that is the case, then what is the "real prose" a student

produces and how can we ever find out what it is, let alone assess it? I hasten to add that this is only my impression as a tester and teacher of writing. I have not conducted any formal research on this topic, nor do I know of anyone who has. But the topic is important, it seems to me, and one to which we should seek an answer. (Note that I have avoided an even more important and difficult question: what is good writing?)

It is also a tenet of faith among writing teachers that multiple-choice tests cannot assess writing. The research literature on this question is mixed. Usually the question asked is for what purpose is writing being assessed? For placement into writing courses? For exemption from writing courses? To identify exceptionally good or bad writing? To diagnose writing problems? Depending on the purpose of the test, a multiple-choice test may very well be appropriate. Hunter Breland and Judith Gaynor's study on the Test of Standard Written English (Journal of Educational Measurement, Summer, 1979) concludes that the results of that test can be used to identify students who are weak writers and could benefit from additional instructions in writing. Other studies reach a variety of conclusions both for and against the use of multiple-choice tests to assess writing. All of the research, I would suggest, misses the basic question: what aspects of writing can multiple-choice items assess? Moreover, what aspects of writing can different item types

assess? How important are item types? Which item types work and which do not? There is little research on item types and their relation to the assessment of writing.

A careful reading of much of the literature on writing assessment reveals that it is the product of questionable research methodology and procedures. It is not uncommon to read an article which arrives at a quite definite conclusion (the quality of student handwriting does affect reader rating of essays, for example), but the conclusion is based upon research conducted in one English Department using student essays selected from among freshman composition classes and scored by teaching assistants with questionable training.

Perhaps the greatest weakness in all research in writing assessment is that there is no common, standard, accepted procedure for evaluating writing. As E.D. Hirsch points out in his book The Philosophy of Composition (University of Chicago Press, 1977), the single, most important issue in composition research is to determine a valid, reliable, and accepted method for evaluating writing, for without it all research into composition is impossible. Unless there is a method for determining the quality of writing, how can there be research into the teaching of writing and the assessment of writing? Some researchers use primary trait scoring, others use holistic scoring, while a few use analytic scoring. And there is widespread disagreement on the validity of each

method. As long as such disagreement exists, and as long as there is no commonly accepted method or methods for evaluating writing, there will be little progress in research on the assessment of writing.

Let me illustrate this problem more specifically. I am at present conducting a study based upon the results of the first year of testing in New Jersey. Using the results for 42,982 entering freshmen in 2 and 4-year colleges in New Jersey, I am correlating their scores on a 125-item multiple-choice test of reading and writing with their scores on a 20-minute essay. The essay was scored holistically using a 4-point scale in a controlled reading conducted by ETS. I am also correlating scores on various configurations of the multiple-choice test with the essay score. For example, I have selected certain item types and have correlated the score on a group of items of a specific type with the essay score. The preliminary results are most interesting and have confirmed some of my beliefs, dashed others, and raised even more questions. I hope to complete this study in the coming months. But I would point out that this is a limited study designed to answer only a few of the basic questions I mentioned earlier. Also, the results of this study depend upon the use of holistic scoring to evaluate writing. But my study is based upon extensive data gathered statewide from a great variety of students. Also, the multiple-choice test used in the testing program was designed and written by ETS in cooperation with many college

teachers of reading and writing in New Jersey. The test was extensively pre-tested and refined. The holistic scoring was carefully controlled and conducted using experienced teachers. In short, the data I am using were gathered from a very expensive and extensive testing program.

In cooperation with some ETS staff members, I have designed another study to answer some of the questions I raised earlier in this paper. The study would gather writing samples from both college and high school students. The samples would consist of both 20 and 45-minute essays written both in and out of class over a period of time. There would also be multiple-choice tests using carefully designed item types. The study would attempt to answer some of the other questions I raised earlier in my paper. Unfortunately, the study exists only on paper since there is no funding for it.

This raises the final issue in the assessment of writing. Just as the proper assessment of writing is very expensive, so too is research into the assessment of writing very expensive. As we all know, testing costs money, and research into testing costs even more money. In addition to all the problems of research into the assessment of writing which I have mentioned here, I will close by mentioning the final problem---lack of adequate funding to seek the answers to questions which need to be answered.

William Lutz
English Department
Rutgers University
Camden, NJ 08102

What We Know and Don't Know: Needed Research in Writing Assessment

William Lutz

Abstract

After an extensive review of the available research on large-scale writing assessment, certain issues in writing assessment seem to be unresolved, and still other issues are not supported by adequate research. This paper reviews the basic issues in writing assessment, points out which topics are supported by strong research, and which topics are not supported by research and which need to be studied further before any definitive statements can be made about them. Among the topics discussed as needing research are the use of the 45-minute essay as opposed to the 20-minute essay, the use of multiple-choice tests to assess writing, and the effect of item types on writing assessment. Specific topics for further research are listed.